

PREPARED FOR SUBMISSION TO JHEP
Cavendish-HEP-16/16,
KEK-TH-1936,
PITT-PACC 1609

Quark-gluon discrimination in the search for gluino pair production at the LHC

Biplob Bhattacharjee,^a Satyanarayan Mukhopadhyay,^b Mihoko M. Nojiri,^{c,d}
Yasuhito Sakaki^e and Bryan R. Webber^f

^a*Centre for High Energy Physics, Indian Institute of Science, Bangalore, India*

^b*PITT-PACC, Department of Physics and Astronomy, University of Pittsburgh, PA 15260, USA*

^c*Kavli IPMU (WPI), The University of Tokyo, Kashiwa, Chiba 277-8583, Japan*

^d*KEK Theory Center and Sokendai, Tsukuba, Ibaraki 305-0801, Japan*

^e*Department of Physics, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea*

^f*Cavendish Laboratory, J.J. Thomson Avenue, Cambridge, UK*

E-mail: biplob@cts.iisc.ernet.in, satya@pitt.edu,
nojiri@post.kek.jp, sakakiy@post.kek.jp,
webber@hep.phy.cam.ac.uk

ABSTRACT: We study the impact of including quark- and gluon-initiated jet discrimination in the search for strongly interacting supersymmetric particles at the LHC. Taking the example of gluino pair production, considerable improvement is observed in the LHC search reach on including the jet substructure observables to the standard kinematic variables within a multivariate analysis. In particular, quark and gluon jet separation has higher impact in the region of intermediate mass-gap between the gluino and the lightest neutralino, as the difference between the signal and the standard model background kinematic distributions is reduced in this region. We also compare the predictions from different Monte Carlo event generators to estimate the uncertainty originating from the modelling of the parton shower and hadronization processes.

Contents

1	Introduction	1
2	Analysis Setup	3
2.1	Overview of quark-gluon tagging variables	3
2.2	Kinematic selection of signal region	4
2.3	Monte Carlo simulation of signal and background processes	5
3	Results	7
3.1	MC truth level quark-gluon fraction	7
3.2	Inclusive and exclusive kinematic variables	8
3.3	Multivariate analysis	10
3.4	Projected reach in $M_{\tilde{g}} - M_{\tilde{\chi}_1^0}$ plane	13
4	Summary and Outlook	14

1 Introduction

The current LHC search for strongly interacting supersymmetric particles in a multi-jet final state primarily relies on kinematic discriminants to separate the signal from very large standard model (SM) backgrounds [1–4]. The signal from heavy squarks or gluinos decaying to a light neutralino lies in the high visible and missing momentum tail. The hadronic jets in the supersymmetry (SUSY) signal come from the decay of a gluino or a squark to one or more quarks and a neutralino. On the other hand, for the dominant SM background of a single weak boson and multiple jets, the jets originate from the initial state radiation of both quarks and gluons. A natural question therefore is whether the difference in the substructure of the decay jets in the signal process and the radiation jets in the SM background can be utilized to further improve the searches. This difference is related to the discrimination of quark- and gluon-initiated jets [5–16], a topic being actively explored by the ATLAS and CMS collaborations [17–19].

One goal of the studies in this direction by ATLAS and CMS is to derive template distributions from data for observables that can separate quark- and gluon-initiated jets [17–19]. Such a data-driven approach can avoid uncertainties coming from the Monte Carlo (MC) modelling of the low energy hadronization component, and to a lesser extent of the parton shower implementation. Although the data based templates are still in an early stage of development, especially when employing multiple

observables, which requires large statistics not yet reached at the LHC, it is worthwhile to briefly discuss the method and its comparison to MC predictions. The only input from MCs in this approach is the quark and gluon-initiated jet fractions in two different processes ¹, for example, in the dijet and γ +jet or Z +jet process, computed at the Born level including parton shower effects. With this definition, depending upon the jet transverse momenta, the dijet event sample consists of 50 – 60% gluon-initiated jets, while the γ +jet or $Z(\rightarrow e^+e^-)$ +jet events contain 70 – 80% quark-initiated jets. The observed normalized distribution in the data for a given observable in the two samples can then be used to derive the normalized distribution for a “pure” quark and gluon jet by solving a pair of linear equations in two variables (for each bin of the normalized distribution, and repeated for different transverse momenta and rapidity intervals). While the uncertainties coming from the parton distribution functions and MC implementation of the Born process and initial and final state radiation are small, the largest systematic uncertainty for such studies arises from the dependence of the templates on the processes being used to derive it [19]. This key aspect of process dependence requires further studies before data-driven templates can be employed in the long run for physics searches without bringing in additional large systematics.

Comparison of the templates derived from data and from the MCs following the same procedure described above shows that while the MC predictions for quark-initiated jets agree reasonably well with the distributions extracted from data, the distributions for the gluon-initiated jets differ [17–19]. The data based templates for gluon jets fall in most cases in between the predictions of different MCs (*Pythia* [20, 21] and *Herwig* [22] to be specific). Such difference in MC prediction for the distribution of quark-gluon tagging observables have also been observed in phenomenological studies [8–14, 16]. With this in mind, the usefulness of quark-gluon discrimination in physics searches at the LHC can be studied using existing event generators, and future use of data-driven templates is expected to lead to a performance somewhere in between the *Pythia*- and *Herwig*-based predictions. If promising improvements are found irrespective of the MC used, and after folding in additional systematic uncertainties in the background prediction from substructure variables, the physics case to pursue quark-gluon discrimination as a tool to find new physics at the LHC would be well motivated.

The goal of this study is to evaluate the expected improvement in the search for gluino pair production at the LHC by including the quark-gluon tagging observables to the standard supersymmetry search strategy in the multijet and missing transverse momentum channel. After including initial and final state parton shower effects to leading order matrix elements, it is estimated that while the third and fourth highest transverse momentum jets in gluino-pair events are expected to be quark-initiated,

¹If more than two processes can be used, one has a cross-check on the results [19].

in the dominant V +jets ($V = Z, W$) backgrounds, they are more likely to be gluon-initiated. This leads to a considerable improvement in the signal to background ratio, when jet substructure based observables are utilized. Moreover, including both the kinematic and the jet substructure observables within a multivariate analysis is found to enhance the search prospects further, especially when the mass difference between the gluino and the neutralino lies in an intermediate region. The projected improvement over standard kinematics based searches is observed independent of the MC generator used, though to a different degree.

In Sec. 2, we describe the quark-gluon separation variables used to define a multivariate discriminant, our Monte Carlo simulation of the signal and background processes as well as the kinematic selection of the signal region. We begin Sec. 3 by first describing the expected quark-gluon fraction of jets in the signal and background processes based on truth level MC information. This is followed by a discussion on the distribution of relevant kinematic variables. The multivariate analysis procedure is described next, followed by the results on the boosted decision tree based separation of the signal and background jet substructure. Combining the information from both kinematics and jet substructure we obtain the signal and background likelihood distributions, which are then used to estimate the expected LHC search reach using different methods in the gluino-neutralino mass plane. We summarize our findings in Sec. 4.

2 Analysis Setup

2.1 Overview of quark-gluon tagging variables

Based on the difference in splitting probabilities in a parton-shower picture, different possible variables have been proposed for quark-gluon discrimination, which essentially rely on the fact that a gluon produced with similar kinematics leads to a larger multiplicity of soft emissions compared to a quark, and a gluon-initiated jet is wider than a quark-initiated one. These differences follow from the higher colour charge-squared of the gluon, $C_A = 3$, versus $C_F = 4/3$ for a quark. As demonstrated in previous studies, based on both perturbative methods as well as MC simulations, it is found that the following variables lead to a better quark-gluon separation:

1. The number of charged tracks inside the jet cone (n_{ch}), with each charged track having $p_T > 1$ GeV, where p_T denotes its transverse momentum. Even though it is difficult to model this observable accurately by MC generators, the recent ATLAS studies on the charged track multiplicity distribution using 8 TeV LHC data shows reasonable agreement for a set of MC tunes upto very high jet transverse momenta [23]. We shall utilize such tunes in our study for both Pythia and Herwig MCs, as discussed in Sec. 2.3.

2. Energy-energy-correlation (EEC) angularity [9] variables, for example, the observable denoted by $C_1^{(\beta)}$ can be defined in terms of the charged track momenta as

$$C_1^{(\beta)} = \frac{\sum_i \sum_{j>i} p_{T,i} \times p_{T,j} \times (\Delta R(i,j))^\beta}{(\sum_i p_{T,i})^2}. \quad (2.1)$$

Here, the sums over i and j run over all the tracks associated to the jet with $j > i$, while β is a tunable parameter. As determined in previous studies [9], from perturbative calculations and MC simulations, $\beta = 0.2$ is found to be an optimal choice that maximizes the quark-gluon separation. The distance in the rapidity-azimuthal angle plane between the tracks i and j is denoted by $\Delta R(i,j)$.

3. Jet mass (m_J) scaled by its transverse momentum $m_J/p_{T,J}$.
4. In addition to the above set of variables, as discussed in our previous study [10], the input for the number of softer reconstructed jets (associated jets) around a primary hard jet can also improve quark-gluon separation, since it captures additional information from radiation outside the jet radius not included in the above variables.

In this study we shall use n_{ch} , $C_1^{(\beta)}$ and $m_J/p_{T,J}$ as the inputs to a multivariate discriminant for quark- and gluon-like jets. While the inclusion of associated jets can be helpful, it is challenging to do so in a multijet environment, as one needs to remove overlap with ISR jets. We leave the investigation of such overlap removal methods to a future study.

2.2 Kinematic selection of signal region

The ATLAS and CMS searches define multiple signal regions determined in terms of kinematic selection criteria that can separate a SUSY squark or gluino production process from the SM backgrounds in the multijets+ \cancel{E}_T channel. Even though this is a challenging analysis in an hadronic environment, for high squark-gluino masses the hard scale of the signal process is higher than the hard scale of most SM processes. This latter fact is reflected in the high values of sum of jet transverse momenta (H_T) or effective mass ($M_{\text{eff}} = H_T + \cancel{E}_T$) demanded in the signal regions. Following the ATLAS search strategies for 14 TeV LHC [24], we first make a pre-selection of events based on the following cuts:

Cut-1:

1. The number of jets, $n_j \geq 4$, with $p_T^{j_1} \geq 160$ GeV and $p_T^{j_2, j_3, j_4} \geq 60$ GeV. For all other jets we demand $p_T^j \geq 20$ GeV. The rapidity coverage of the jets is determined by ATLAS calorimeter design, where the forward calorimeter covers the pseudo-rapidity range of $|\eta| < 4.9$. However, the tracker covers only

upto $|\eta| < 2.5$, and therefore it is not possible to obtain the information on the number of charged tracks inside jets in the forward region. Since the quark-gluon discrimination variables can be more accurately determined in terms of charged track momenta, we therefore count n_j only within $|\eta| < 2.5$.

2. No isolated lepton (electron or muon) with $p_T > 10$ GeV, within $|\eta| < 2.5$.
3. Missing transverse momentum in the event $\cancel{E}_T > 160$ GeV.
4. $\Delta\phi(\text{jet}, \cancel{E}_T)_{\min} > 0.4$ (0.2) radian for j_1, j_2, j_3 (for all other jets with $p_T > 40$ GeV).

The jet p_T cuts and the \cancel{E}_T cut are applied at the matrix element (ME) level while generating the background events, which is modelled by $Z(\rightarrow \nu\bar{\nu})+\text{jets}$. Furthermore, in order to obtain a large statistics of events with a high M_{eff} cut, we have generated several different samples of the $Z+\text{jets}$ events, one with each value of the M_{eff} cut. As discussed in detail in Sec. 2.3, we normalize our total $Z+\text{jets}$ event rate by comparison with the number of events reported in the ATLAS simulation after the cuts in 4jm category [24] (defined as **Cut-1** followed by $\cancel{E}_T/M_{\text{eff}} > 0.25$ and $M_{\text{eff}} > 3200$ GeV). With this, we are able to reproduce with a reasonable accuracy the ATLAS projected sensitivity in the $M_{\tilde{g}} - M_{\tilde{\chi}_1^0}$ plane for 14 TeV LHC with 300 fb^{-1} data.

In addition to the above basic set of cuts, in order to compare with the search reach of ATLAS 14 TeV projections [24], we have computed the signal and background event yields in seven different signal regions (4j1, 4jm, 4jt, 5j, 6j1, 6jm, 6jt) as defined in the ATLAS study [24], essentially differing in the values of the M_{eff} , $\cancel{E}_T/M_{\text{eff}}$ and $\cancel{E}_T/\sqrt{H_T}$ cuts.

2.3 Monte Carlo simulation of signal and background processes

For both the signal and background processes, the parton level matrix elements are computed, and the events generated using **MG5aMC@NLO** [25]. The parton level events are passed onto both **Pythia** 6.4.28 (with the P2012-RadLo tune) [20], and **Herwig++** 2.7.1 (with the default tune) [22], for simulating parton shower, hadronization and underlying events. The above choice for the **Pythia** tune is based on better data-model agreement in a recent ATLAS study comparing the charged track multiplicity distribution in the data with MC predictions [23]. The parton shower and hadronization effects are simulated using two different MCs to estimate the uncertainty in quark-gluon tagging coming from MC modelling. The signal cross-section is normalized to predictions including the resummation of soft-gluon emission at next-to-leading logarithmic accuracy, matched to next-to-leading order supersymmetric QCD corrections [26].

We use the **CTEQ6L1** [27] parton distribution functions from the **LHAPDF** [28] library, and the factorization and renormalization scales are kept at the default

event-by-event choice of `MG5aMC@NLO`. Detector effects have been simulated using `Delphes3` [29], where the jet clustering is performed with `FastJet3` [30]. Jets are reconstructed using the anti- k_T clustering algorithm [30, 31] with radius parameter $R = 0.4$. We have implemented the variables used for studying quark and gluon jet tagging in the `Delphes3` framework.

As the signal process, we consider gluino pair production, followed by its three-body decay with 100% branching ratio to a pair of quarks and the lightest neutralino, via intermediate off-shell squarks. In general, depending on the squark mass, on-shell squark production will also contribute to the same final state. However, for studying the usefulness of quark-gluon tagging in SUSY searches, a simplified model with only the gluino and the lightest (bino-like) neutralino is adequate, and the rest of the MSSM particles are assumed to be decoupled. The final state of interest will then be ≥ 4 -jets and missing transverse momentum.

It is well-understood that the primary background to such a multi-jet and missing momentum search comes from Z +jets production (with Z decaying to neutrinos), followed by a similar contribution from W +jets (where the charged lepton from the W boson decay falls outside the tracker acceptance, and therefore is not reconstructed as a lepton). The fractional contribution of $t\bar{t}$ +jets and single top production is reduced at higher M_{eff} regions, but it can also become comparable to the individual weak-boson contributions depending upon the signal region of interest. A strong cut on the \cancel{E}_T variable reduces the QCD multijet background, especially by ensuring that the jet direction and the \cancel{E}_T vector direction are not correlated. For a comparison of different SM background contributions, see, for example, the recent ATLAS note on squark-gluino search at the 13 TeV LHC with 13.3 fb^{-1} of data [2]. Both the recent 13 TeV ATLAS analysis and the ATLAS projection results for 14 TeV LHC with 300 fb^{-1} data show that the total SM background in our signal region of interest (i.e., after **Cut-1** and with $M_{\text{eff}} > 1.8 \text{ TeV}$) is always less than twice the Z +jets contribution. The kinematic and quark-gluon fraction properties in Z +jets and the subdominant W +jets processes are nearly identical. Therefore, we perform the MC simulations using only the Z +jets process, and take the total SM background as twice the Z +jets prediction, which is a conservative estimate.

Since we shall focus on a multivariate analysis (MVA) strategy especially for the quark-gluon separation, the statistics of MC events required to perform the boosted decision tree (BDT) training is very high, especially if the number of input variables to the BDT training is large (eventually we shall use a ten variable BDT). Furthermore, these event samples are all required to pass a pre-selection of **Cut-1** and different values of high M_{eff} cuts. Therefore, generating such a large statistics of events with matrix element (ME) - parton shower (PS) matching is beyond the scope of our computational resources. On the other hand, as is well-known, to obtain accurate predictions for the jet p_T s in processes such as Z +jets, ME-PS matching is important. However, since we are primarily interested in four relatively hard and

central jets, the expectation is that events based on $Z + 3$ -jets or $Z + 4$ -jets matrix elements followed by PS can cover the relevant phase space region, and therefore the normalized differential distributions should be well-predicted by these event samples. In order to check this fact, we generated three different samples of Z +jets events and compared all the kinematic and jet-substructure distributions between them. The three samples are: (1) Z +jets, ME-PS merged upto 4-jets, (2) $Z + 3$ -jet ME followed by PS and (3) $Z + 4$ -jet ME followed by PS. We find that all the distributions have very similar shape in the three samples (as shown in the Appendix). Thus it is possible to obtain accurate normalized distributions by just using the $Z + 3$ -jet ME (followed by PS) event sample, for which generating a large enough statistics is least resource intensive among the three. For the overall normalization, as discussed earlier, we normalize our Z +jets event yield to the number reported in ATLAS simulation [24], and take the total SM background as two times the Z +jets contribution.

3 Results

3.1 MC truth level quark-gluon fraction

As discussed in Sec. 2.3, in the signal process of gluino pair production, with gluino dominantly decaying via (onshell or offshell) squarks, the decay jets are all quark-initiated. In addition, there are additional jets in the signal events from initial state radiation (ISR), which may reduce the difference between the signal and background likelihoods if a gluon-initiated ISR jet is harder than the decay jets and also lies in the central region of the detector. At Born level, the dominant background of Z +jets has a higher gluon fraction in the third and fourth highest p_T jets (denoted by j_3 and j_4 respectively). It is thus expected that the maximum discriminating power in the likelihood would come from j_3 and j_4 , rather than the first and second highest p_T jets (denoted by j_1 and j_2).

To define the MC truth level quark and gluon jet fraction, we adopt the following method. Assume that we are looking for quark jets in an event. In the first step we find quarks in the matrix element, and a quark of flavour f is denoted by f_i . Next, in the parton history related to the mother parton i , we find the parton P_i with the same flavour as f_i (we choose the parton with the highest transverse momentum if there are multiple quark partons of flavour f). Finally, if the distance between the jet J and the parton P_i is less than the jet cone size, $\Delta R(J, P_i) < R = 0.4$, we define the jet J as a quark jet. If not, then J is defined as a gluon jet. We emphasize that in the actual study of signal-background discrimination, this definition does not play any role, since in that case, we compare the likelihood of an event being signal-like or background-like, based on an MVA with the discriminating variables as inputs.

For illustration, we show in Tab. 1 the parton level quark fraction of the first four jets, as defined above. A representative signal point with $M_{\tilde{g}} = 2000$ GeV and

$M_{\tilde{\chi}_1^0} = 1000$ GeV has been chosen for Tab. 1, and the quark fractions are shown after the preselection of **Cut-1** and with $M_{\text{eff}} > 1.8$ TeV. The parton shower MC used for this figure is **Pythia 6.4.28**. In general, we see from this table that among the first four hardest jets, most signal events contain 3 – 4 quark jets, while most Z +jets events contain 1 – 2 quark jets.

Process	$f_q^{j_1}$	$f_q^{j_2}$	$f_q^{j_3}$	$f_q^{j_4}$
$\tilde{g}\tilde{g}$ +jets	0.92	0.87	0.77	0.64
Z +jets	0.64	0.55	0.27	0.16

Table 1. Quark fraction (f_q) at the MC truth level for the first four highest- p_T jets in $\tilde{g}\tilde{g}$ +jets and Z +jets processes. All events are selected after passing the jet- p_T , \cancel{E}_T (**Cut-1**) and $M_{\text{eff}} > 1.8$ TeV cuts, at the 14 TeV LHC. See text for details on the determination of f_q .

3.2 Inclusive and exclusive kinematic variables

In the dominant SM background processes of Z/W +jets, the jets come from initial state QCD radiation, which exhibits a strong ordering of the jet p_T s for a given H_T value, primarily because of the enhancement in the soft gluon emission probability given by the QCD splitting functions. On the other hand, for the decay jets coming from gluino decay, the jet transverse momenta are not in general so strongly ordered, as in this case the p_T s of the jets are determined by the mass-gap between the gluino and the lightest neutralino and the mass of the lightest neutralino itself. Admittedly, this is then a SUSY parameter dependent statement as to how the ordered jet p_T distributions would differ between the signal and the background. Nevertheless, for certain ranges of the gluino and neutralino masses, the transverse momentum of the first four jets, ordered according to their p_T s, can carry additional information not entirely captured in the M_{eff} or H_T distributions. We use the nomenclature of exclusive kinematic variables to refer to the ordered jet p_T s, while we shall refer to M_{eff} , H_T and \cancel{E}_T as inclusive kinematic variables.

We show in Fig. 1, the normalized (to unit area) distributions for the kinematic variables used as inputs in defining the combined signal and background likelihood functions, after the event pre-selection of **Cut-1** and an M_{eff} cut of 1.8 TeV. For the signal, we show the distributions at a benchmark point with $M_{\tilde{g}} = 2000$ GeV and $M_{\tilde{\chi}_1^0} = 1000$ GeV, and for illustration results from only the **Pythia** MC are presented. Since only events passing **Cut-1** and $M_{\text{eff}} > 1.8$ TeV are included, the H_T and p_{Tj} distributions have a non-standard shape (first rise to a peak value and then fall). As we can see from this figure, for this signal benchmark point, the exclusive kinematic variables also provide discriminating power over the Z +jets background. For the gluino pair production events, we have also checked that including additional jets in

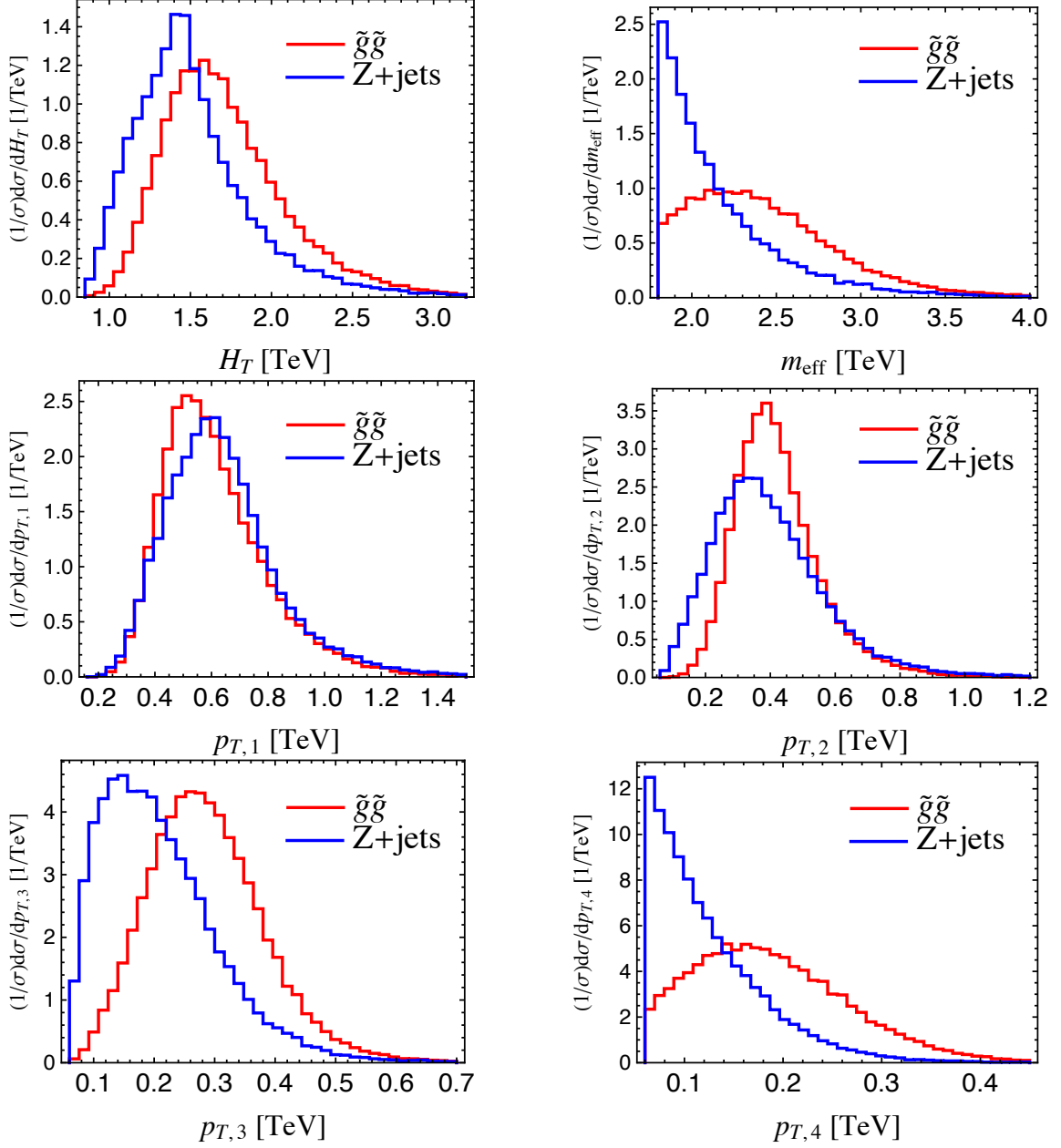


Figure 1. Normalized distribution of inclusive and exclusive kinematic variables. For the signal, we show the distributions at a benchmark point with $M_{\tilde{g}} = 2000$ GeV and $M_{\tilde{\chi}_1^0} = 1000$ GeV. The distributions are presented after Cut-1 and an M_{eff} cut of 1.8 TeV, and for this reason the H_T and $p_{T,j}$ distributions have a non-standard shape.

the matrix element and using ME-PS matching, the kinematic distributions do not show any significant difference.

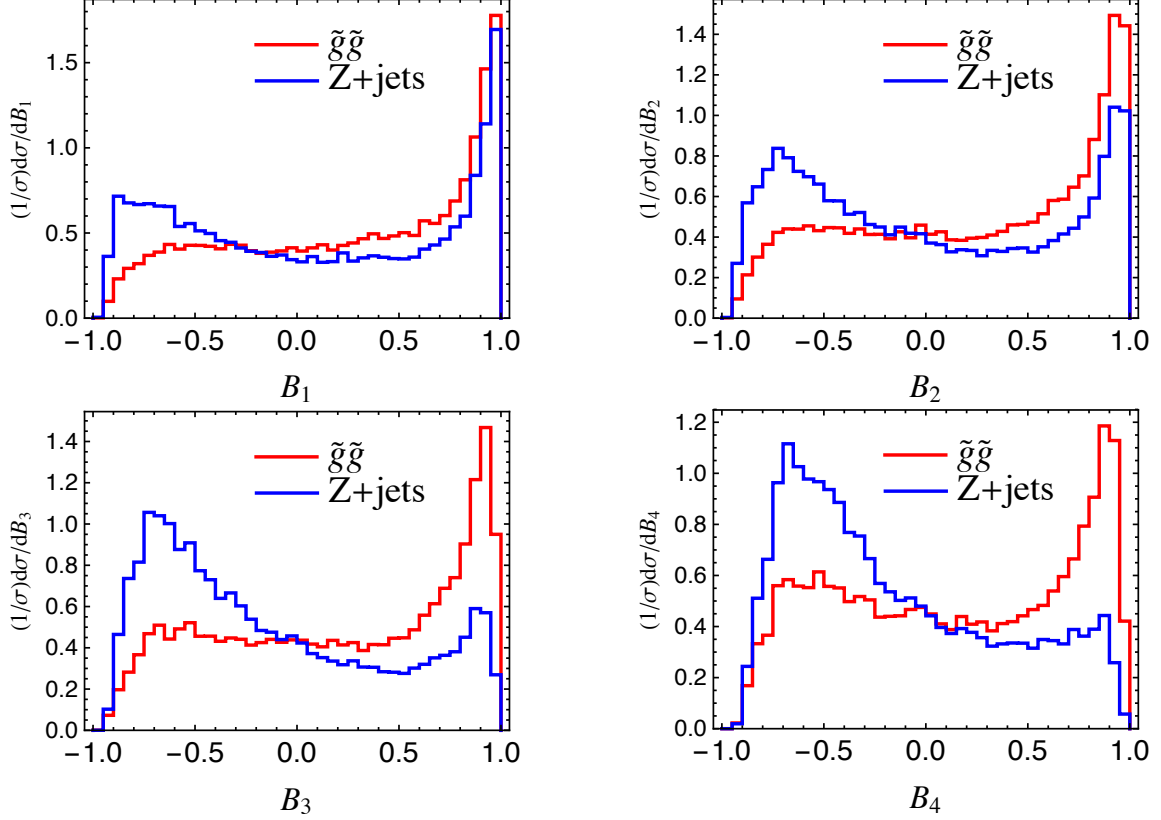


Figure 2. Normalized BDT score distributions based on quark-gluon tagging variables of individual jets ordered according to their p_T (B_i refers to the BDT score for j_i). For illustration, the distributions are shown for a signal benchmark point of $M_{\tilde{g}} = 2000$ GeV and $M_{\tilde{\chi}_1^0} = 1000$ GeV, and using the Pythia6 MC.

3.3 Multivariate analysis

Using the quark-gluon separation variables described in Sec. 2.1, namely, n_{ch} , $C_1^{(\beta)}$ and $m_J/p_{T,J}$ as inputs, we first develop an optimized discriminant using a multivariate analysis. This has been carried out by employing a Boosted Decision Tree (BDT) algorithm with the help of the TMVA-Toolkit [32] in the ROOT framework [33]. The training of the BDT classifier has been performed using the $Z+q$ and $Z+g$ processes at the Born level. The MC samples for these processes are generated such that we obtain an uniform statistical coverage across the entire jet p_T range of interest, and the BDT training is performed for different p_T ranges taken as different categories.

Following the above method, for the signal and background processes, we compute the BDT score B_i for each of the first four jets ordered according to their p_T . This procedure has been carried out using both the Pythia6 and Herwig++ MCs to simulate the parton shower and hadronization aspects. In Fig. 2, we show the distribution of the BDT scores for the first four highest p_T jets in the gluino pair signal and the Z +jets background processes (for illustration, the distributions are shown

using Pythia6). As expected from the truth level quark-gluon fractions discussed in Sec. 3.1, significant separation in the BDT scores for the third and fourth highest p_T jets (B_3 and B_4) are observed, for which the signal jets are mostly quark-initiated, and the background ones are mostly gluon-initiated.

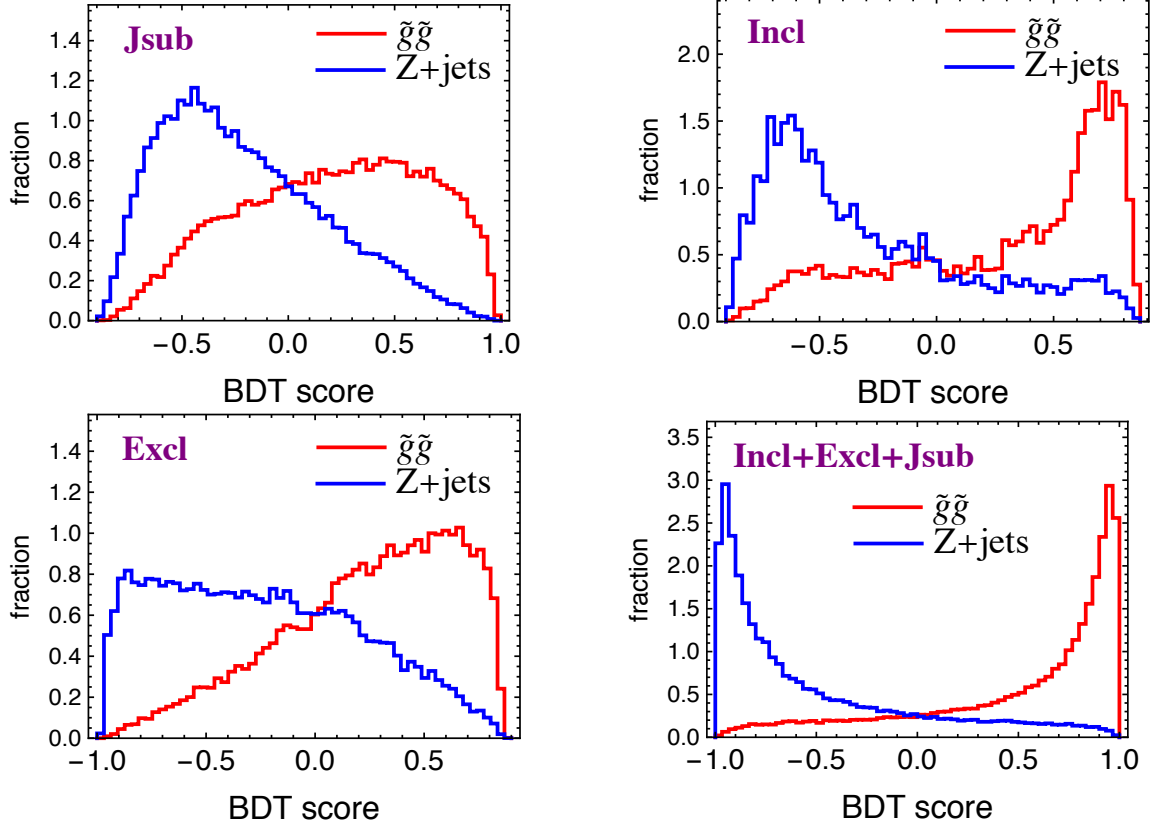


Figure 3. Signal and background likelihood distributions, using as inputs different subsets of variables. Incl., Excl. and Jsub refer to MVA with the input sets $\{M_{\text{eff}}, H_T\}$, $\{p_{T,j1}, p_{T,j2}, p_{T,j3}, p_{T,j4}\}$, and $\{B_1, B_2, B_3, B_4\}$ respectively. The bottom right plot is obtained using an MVA with all ten variables as inputs.

As a final ingredient to our analysis, we perform a further MVA study with ten input variables containing: $\{M_{\text{eff}}, H_T, p_{T,j1}, p_{T,j2}, p_{T,j3}, p_{T,j4}, B_1, B_2, B_3, B_4\}$. This defines a signal and background likelihood with all the kinematic and jet substructure information of the event. The BDT score cut is chosen to maximize the exclusion (or discovery) significance for a given model point. For illustrating the separation power from each subset of variables, we show in Fig. 3 the BDT score distributions obtained with the inclusive kinematic variables (M_{eff} and H_T), the exclusive kinematic variables ($p_{T,j1}, p_{T,j2}, p_{T,j3}$ and $p_{T,j4}$), and the jet substructure based BDT variables (B_1, B_2, B_3 and B_4). We also show in the bottom right panel of Fig. 3 the signal-background separation with all ten variables included together in the MVA. The results are shown for the signal point ($M_{\tilde{g}} = 2000$ GeV and $M_{\tilde{\chi}_1^0} = 1000$ GeV)

and with Pythia MC.

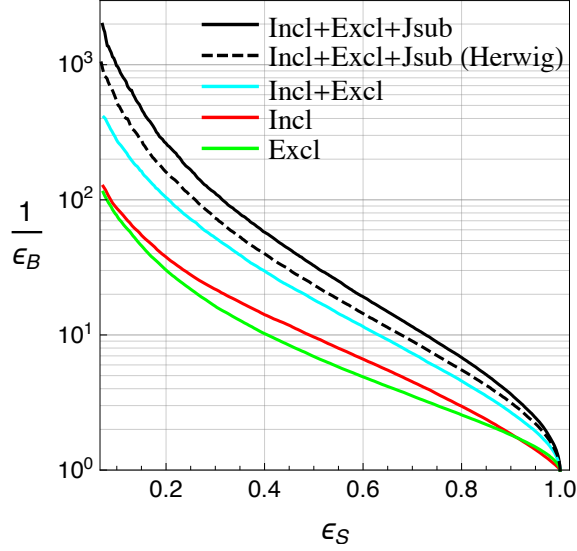


Figure 4. Signal acceptance (ϵ_S) versus inverse of the background acceptance ($1/\epsilon_B$) efficiencies as a function of the BDT cut, for different MVA methods. The solid lines are obtained using Pythia6, while the dashed ROC curve is obtained using the Herwig++ MC. The ROC curves predicted by Herwig++ in the Incl, Excl and Incl+Excl categories are similar to the Pythia6 ones. See text for details.

Based on the final BDT score distribution with ten observables, we can now obtain the ROC curve which shows the signal acceptance (ϵ_S) versus background rejection ($1 - \epsilon_B$) efficiencies as a function of the BDT cut. In Fig. 4, the red, green and cyan curves show the ROC curves for the MVA analyses based on the inclusive, exclusive and inclusive and exclusive variable sets combined respectively. These two sets carry independent information, and therefore the background rejection using the combined set increases compared to the ones using the individual sub-sets, and the ϵ_S and $1/\epsilon_B$ values on the cyan curve is roughly given by the product of their corresponding values on the red and green curves. For example, the efficiencies ($\epsilon_S, \epsilon_B^{-1}$) for the red and the green curves pass through $(0.4, 10)$ and $(0.4, 1.4 \times 10)$ respectively, and the cyan curve passes through the product $(0.4^2, 1.4 \times 10^2)$. The black solid and dashed curves show the performance of the MVA analysis with all the variables taken together, using Pythia6 and Herwig++ respectively. We find that, by adding the jet substructure variables to the MVA, the background rejection factor increases by about a factor of 4 in the Pythia results and by a factor of 2 – 2.5 in the Herwig results, for $\epsilon_S \sim 0.1$. As we shall see in the next subsection, this latter improvement has a considerable impact while considering the exclusion (discovery) reach in the $M_{\tilde{g}} - M_{\tilde{\chi}_1^0}$ plane.

3.4 Projected reach in $M_{\tilde{g}} - M_{\tilde{\chi}_1^0}$ plane

By varying the BDT score cut with all or a subset of observables as input, we can choose the cut that maximizes the significance for a given SUSY parameter point. Here the significance is defined as $S/\sqrt{B + (\delta_{\text{Incl}} + \delta_{\text{Excl}} + \delta_{\text{Jsub}})^2 \times B^2}$, where S and B denote the signal and background event yields for a given integrated luminosity, and δ_{Incl} , δ_{Excl} and δ_{Jsub} are the systematic uncertainties in the background prediction coming from inclusive, exclusive and jet substructure observables respectively. For our significance computation we have set $\delta_{\text{Incl}} = \delta_{\text{Excl}} = \delta_{\text{Jsub}} = 10\%$, and to obtain a conservative estimate of the reach we have added these uncertainties linearly, making the total systematic uncertainty in the background yield prediction to be 30%, when all the variables are included together. We should note that the above definition of the statistical significance may not be accurate when the number of events are small, and especially if S and B are comparable in magnitude. However, we have checked that for most of our parameter points on the exclusion contours, the results do not change appreciably on using a Poisson log-likelihood ratio for the statistical uncertainty component.

In Fig. 5, we show the projected 95% C.L. exclusion contours in the $M_{\tilde{g}} - M_{\tilde{\chi}_1^0}$ plane at the 14 TeV LHC with an integrated luminosity of 300 fb^{-1} . The orange curve is the ATLAS projected sensitivity with standard kinematic cuts (as reproduced by us), while the red, green and blue solid lines show the reach with each subset of variables described in the previous subsection. As expected from the ROC curve in Fig. 4, each of the subsets individually can lead to similar reach in this parameter space. We recall that all the curves include the effect of the pre-selection cuts on the jet p_T 's and $\cancel{E}_T(\text{Cut}-1)$ as well as a high M_{eff} cut. Thus these improvements are within a high mass signal region. It is further observed that on including the information of the ordered jet p_{Ts} of the first four jets the reach improves to a good extent (cyan solid curve). Finally, if we now include the jet substructure information as well, the reach in the $M_{\tilde{g}} - M_{\tilde{\chi}_1^0}$ plane (black solid line) shows considerable improvement over the standard analysis. It should be noted in particular that especially in the region where the mass difference between the gluino and the neutralino falls in an intermediate range, the jet substructure observables provide stronger separation power. We also note that the signal benchmark point used to show the various distributions in this study, namely, $(M_{\tilde{g}} = 2000 \text{ GeV}, M_{\tilde{\chi}_1^0} = 1000 \text{ GeV})$ can be excluded at 2σ level only when the jet substructure variables are included in the MVA. Since we have also included additional systematic uncertainties in the background rate coming from the modelling of both the exclusive and jet substructure observables (upto 30% in total systematic uncertainty), our estimates for the improvement in the LHC reach should be conservative. It is thus promising that utilizing quark-gluon discrimination within an MVA including kinematic observables can considerably improve the LHC search prospects of strongly interacting SUSY particles.

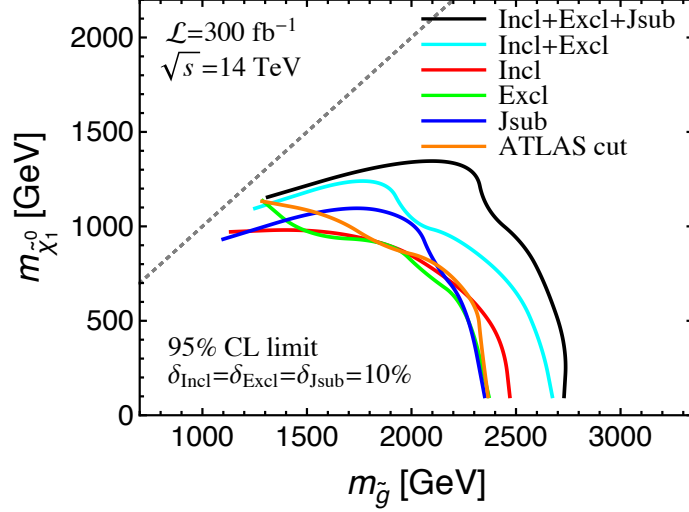


Figure 5. Projected 95% C.L. exclusion contours in the $M_{\tilde{g}} - M_{\tilde{\chi}_1^0}$ plane at the 14 TeV LHC with an integrated luminosity of 300 fb^{-1} . The different systematic uncertainty components have been added linearly, making the total systematic uncertainty in the background yield prediction to be 30%, when all the variables are included together. See text for details on the individual exclusion contours.

In order to understand the uncertainty in the predictions from the MC modelling of jet substructure, we have performed the full analysis using both the `Pythia6` and `Herwig++` MCs. In Fig. 6 we show the 95% C.L. exclusion contours predicted by the two MCs using either only the jet substructure subset (blue curves) or the full variable set (black curves). For reference, the exclusion contours based on ATLAS cuts [24] are also shown (orange curves), and they are almost identical for `Pythia6` and `Herwig++`. The `Pythia6` exclusion contours (solid lines) show a better reach than the `Herwig++` ones (dashed lines), and the difference between the two essentially comes from the jet substructure modelling, which, as remarked earlier, differs significantly for gluon jets. It is however encouraging that both MCs predict significant improvement over the standard analysis. Thus to the extent these two MCs provide an estimate of the uncertainty in prediction, our results show that irrespective of such differences, an improvement is expected in the LHC reach of gluino pair production, especially in the intermediate mass gap region, when we include the quark-gluon separation information within the MVA analysis. Future availability of data-based templates and improved MC tunes are expected to lead to more reliable predictions and a reduction of the systematics in the application of quark-gluon discrimination.

4 Summary and Outlook

Quark-gluon discrimination is becoming a topic of growing interest, both in the theoretical and Monte Carlo front with improved jet substructure based observables being

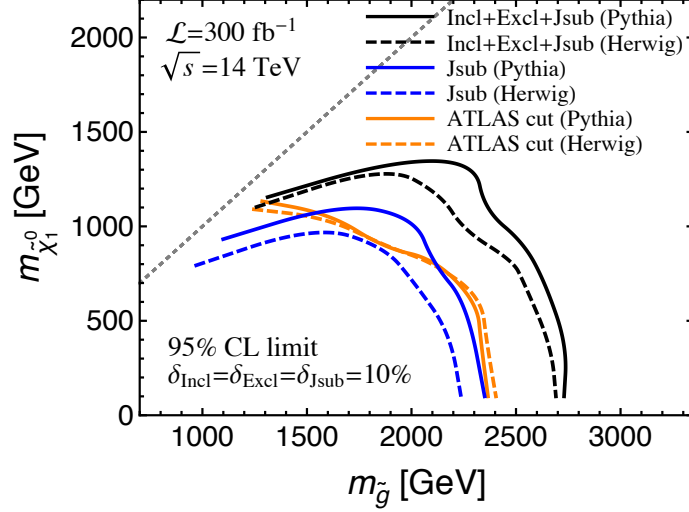


Figure 6. The 95% C.L. exclusion contours predicted by Pythia6 (solid lines) and Herwig++ (dashed lines) using either only the jet substructure subset (blue curves) or the full variable set (black curves). For reference, the exclusion contours based on ATLAS cuts [24] are also shown (orange curves), and they are almost identical for Pythia6 and Herwig++.

designed to capture the detailed pattern of QCD radiation, and on the experimental front with the development of data-based templates for tagging observables as well as validation of existing MC tunes. It is thus an ideal juncture when the importance of quark-gluon jet separation methods in the search for physics beyond the standard model should be thoroughly explored. With this goal in mind, in this paper, we studied the impact of including quark- and gluon-initiated jet discrimination in the search for gluino pair production events at the LHC. As seen in Tab. 1, when ordered according to their transverse momenta, the third and fourth jets are more likely to be quark-initiated for the signal process, while for the dominant background of Z/W +jets, they are more likely to be gluon-initiated. With the quark and gluon separation variables of the number of charged tracks, energy correlation functions C_1^β , and jet mass ($m_J/p_{T,J}$) as inputs to a multivariate analysis, we first develop a BDT-based quark-gluon discriminant across a large range of jet p_T using the $Z + q$ and $Z + g$ processes as the training samples. In addition to the standard “inclusive” kinematic variables of \cancel{E}_T , H_T and M_{eff} , we also observe that for a given H_T value, there is a strong ordering of the jet p_T s for the Z +jets background process, while for the signal process the jets are not so strongly ordered. This is of course a parameter point dependent statement, as the gluino-neutralino mass splitting and the mass of the lightest neutralino determines the ordering of the p_T s of the decay jets. However, in certain regions in the $M_{\tilde{g}} - M_{\tilde{\chi}_1^0}$ plane the inclusion of these “exclusive” kinematic variables within an MVA can help in increasing the signal to background ratio (S/B). We have explored different combinations of the inclusive, exclusive and

jet substructure observables as MVA input variables to understand the importance of each category, and find that all three sub-categories, when added individually to a set of pre-selection cuts and a minimum effective mass cut (chosen according to the working point in the $(M_{\tilde{g}}, M_{\tilde{\chi}_1^0})$ plane), lead to a similar improvement in S/B . Consequently, compared to an optimized kinematic-category based search (as currently carried out by the ATLAS and CMS collaborations), inclusion of the quark-gluon discrimination variables improves the reach in the $M_{\tilde{g}} - M_{\tilde{\chi}_1^0}$ plane, especially in a region where the difference between $M_{\tilde{g}}$ and $M_{\tilde{\chi}_1^0}$ falls in an intermediate range. This is because for such intermediate mass gaps, the H_T and M_{eff} distributions in the signal can become similar to the SM background ones. Given the fact that the jet substructure based variables, as well as the inclusive and exclusive kinematic distributions can bring in additional systematic uncertainties in the background rate determination, we have included a total systematic uncertainty of 30% on our background event yield, which should be a reasonable estimate.

As discussed in the introduction, there exist differences in the Monte Carlo prediction of the quark-gluon separation observables, and the data-based templates for gluon-initiated jets tend to lie in between the predictions of the **Pythia** and **Herwig** MCs, while for quark-initiated jets the data-based templates largely agree with the MCs. With this observation in view, we carry out our complete analysis using both the MC event generators, in order to get an understanding of the variation in signal and background rates from MC modelling of parton shower and hadronization processes. This translates into a variation in the expected reach in the $M_{\tilde{g}} - M_{\tilde{\chi}_1^0}$ plane as well. While the expected improvement in reach does depend upon the MC generator used, the generic patterns remain the same. The reach based on different sets of kinematic variables are similarly predicted by both the event generators, as largely expected, since the low energy hadronization component does not enter in the jet transverse momentum distributions, while the effect of parton shower variation is weaker if we focus on high- p_T jets only. Therefore, the MC variation almost entirely originates from the modelling of the jet substructure. It is however encouraging that independent of the MC generator used, the inclusion of quark-gluon discrimination leads to an improvement in probing the gluino pair production process, especially in the intermediate mass-gap region. This fact, combined with the future prospect of obtaining data-driven multivariate templates that do not rely on the MC modelling of the hadronization component (and possible improvements in the MC tunes as well), makes the utilization of quark-gluon discrimination in new physics searches sufficiently promising. We therefore expect that it would be explored in further detail by the LHC experimental collaborations in the future search for strongly interacting supersymmetric particles.

Appendix

In this appendix, we discuss the details of our simulation of the Z +jets background process. As discussed in Sec. 2.3, due to the necessity to generate several large statistics event samples as an input to the MVA after Cut-1 and different values of high M_{eff} cuts, we use the $Z + 3$ -jets ME (followed by PS) event sample, since it accurately reproduces normalized differential distributions for all the input variables, and is less resource intensive. The overall normalization of the Z +jets background is fixed by comparison with the ATLAS simulation results in the 4jm category [24]. This method is also cross-checked by reproducing to a good accuracy the ATLAS projected exclusion contour [24]. In Figs. 7 and 8, we show the normalized distributions of the inclusive, exclusive and jet substructure variables utilized in this study for the three event samples of (1) Z +jets, ME-PS merged upto 4-jets, (2) $Z + 3$ -jet ME followed by PS and (3) $Z + 4$ -jet ME followed by PS. As we can see from this figure, the difference in shape between these three event samples is negligibly small.

Acknowledgments

YS is grateful to Michihisa Takeuchi for helpful discussions and computing support. The work of BB is supported by the Department of Science and Technology, Government of India, under the Grant Agreement number IFA13-PH-75 (INSPIRE Faculty Award). SM is supported in part by the U.S. Department of Energy under grant No. DE-FG02-95ER40896 and in part by the PITT PACC. M.N. and Y.S. are partially supported by the Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports, and Culture (MEXT), Japan (Nos. 16H06492 and 16H03991 for M. M. Nojiri), and also by the World Premier International Research Center Initiative (WPI Initiative), MEXT, Japan. BW is grateful for the hospitality of Kavli IPMU while part of this work was performed.

References

- [1] G. Aad *et al.* [ATLAS Collaboration], “Summary of the searches for squarks and gluinos using $\sqrt{s} = 8$ TeV pp collisions with the ATLAS experiment at the LHC,” JHEP **1510** (2015) 054.
- [2] The ATLAS collaboration, “Further searches for squarks and gluinos in final states with jets and missing transverse momentum at $\sqrt{s} = 13$ TeV with the ATLAS detector,” ATLAS-CONF-2016-078.
- [3] V. Khachatryan *et al.* [CMS Collaboration], “Searches for Supersymmetry using the M_{T2} Variable in Hadronic Events Produced in pp Collisions at 8 TeV,” JHEP **1505** (2015) 078.

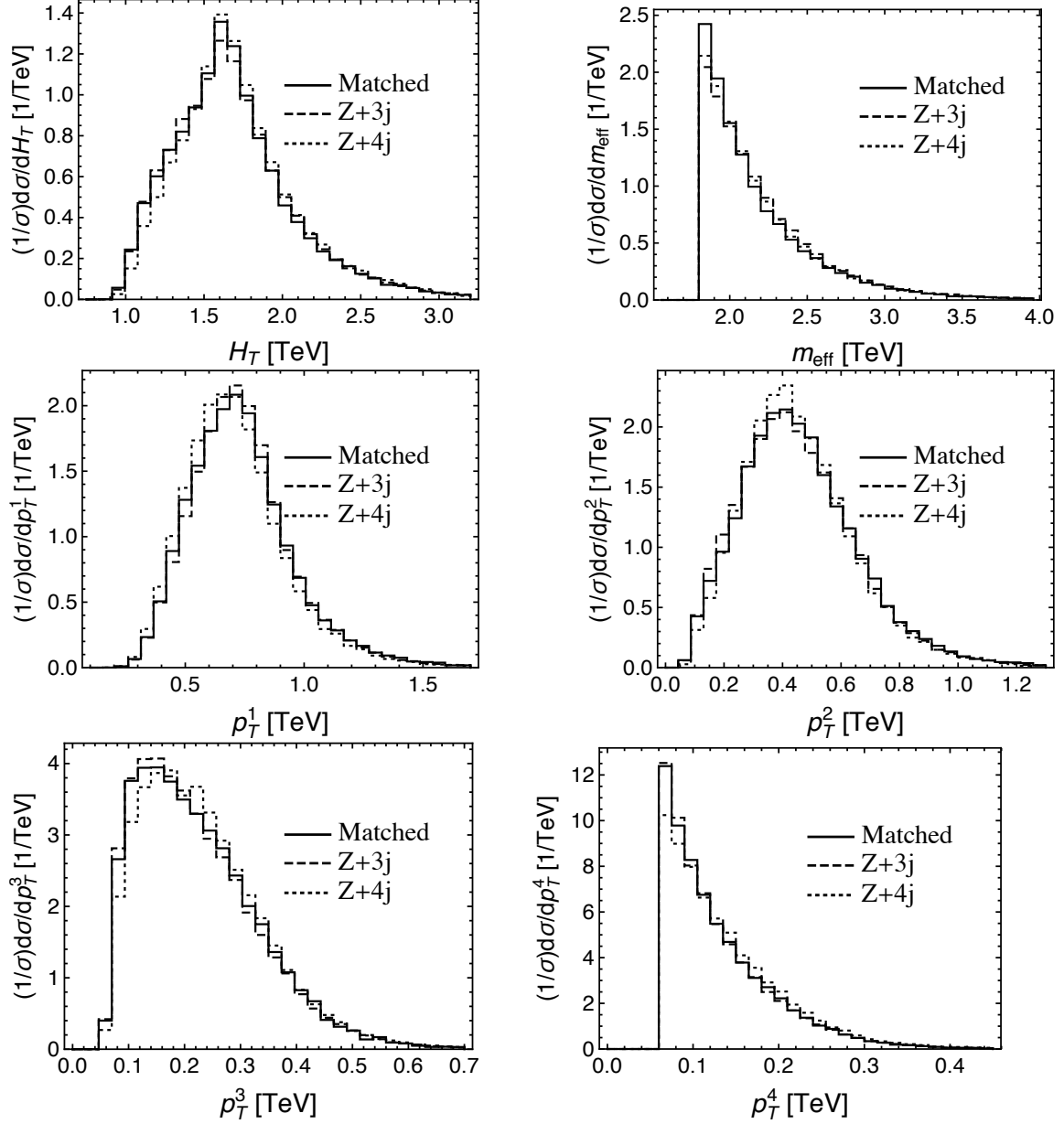


Figure 7. Normalized distribution of the inclusive and exclusive kinematic variables using three different event samples: (1) Z+jets, ME-PS merged upto 4-jets, (2) Z + 3-jet ME followed by PS and (3) Z + 4-jet ME followed by PS. The distributions are presented after Cut-1 and an M_{eff} cut of 1.8 TeV, and for this reason the H_T and p_{Tj} distributions have a non-standard shape.

- [4] CMS Collaboration, “Search for supersymmetry in events with jets and missing transverse momentum in proton-proton collisions at 13 TeV,” CMS-PAS-SUS-16-014.
- [5] J. Gallicchio and M. D. Schwartz, “Quark and Gluon Tagging at the LHC,” Phys. Rev. Lett. **107** (2011) 172001.

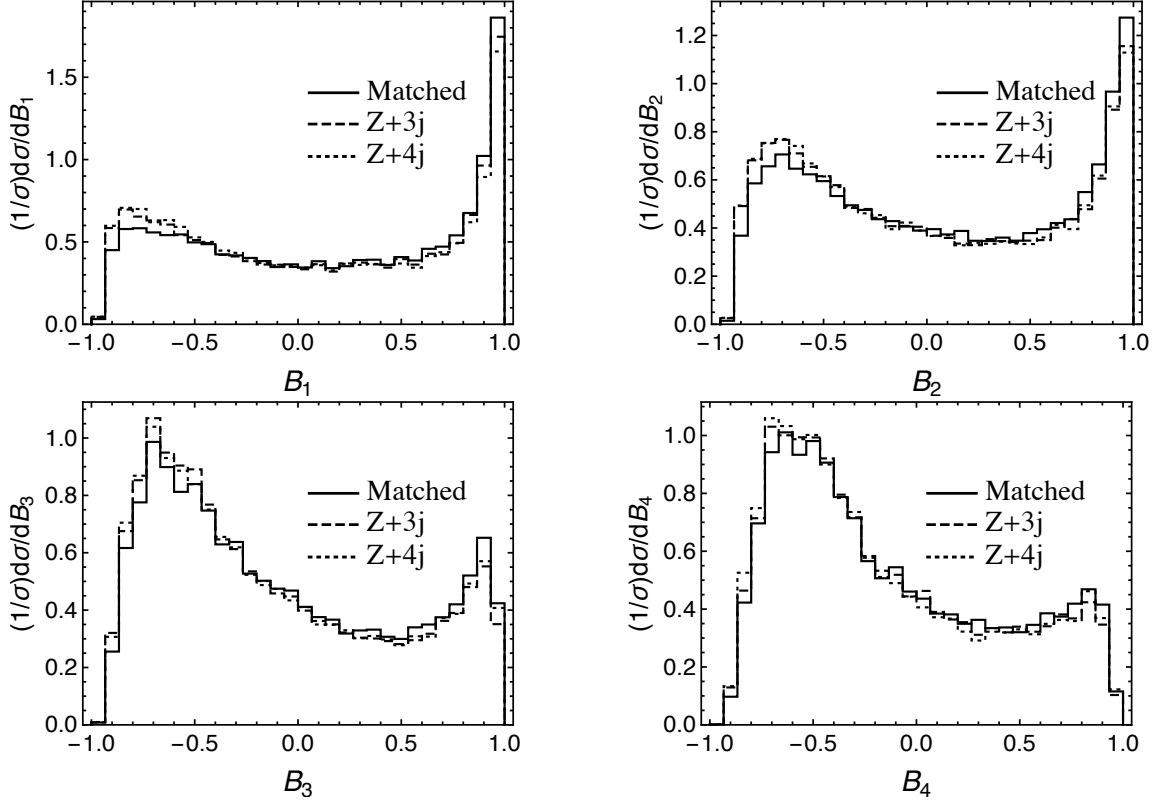


Figure 8. Normalized distribution of the jet substructure based BDT variables, using three different event samples: (1) Z +jets, ME-PS merged upto 4-jets, (2) Z +3-jet ME followed by PS and (3) Z +4-jet ME followed by PS. The distributions are presented after Cut-1 and an M_{eff} cut of 1.8 TeV.

- [6] J. Gallicchio and M. D. Schwartz, “Pure Samples of Quark and Gluon Jets at the LHC,” JHEP **1110** (2011) 103.
- [7] J. Gallicchio, J. Huth, M. Kagan, M. D. Schwartz, K. Black and B. Tweedie, “Multivariate discrimination and the Higgs + W/Z search,” JHEP **1104** (2011) 069.
- [8] J. Gallicchio and M. D. Schwartz, “Quark and Gluon Jet Substructure,” JHEP **1304** (2013) 090.
- [9] A. J. Larkoski, G. P. Salam and J. Thaler, “Energy Correlation Functions for Jet Substructure,” JHEP **1306** (2013) 108.
- [10] B. Bhattacharjee, S. Mukhopadhyay, M. M. Nojiri, Y. Sakaki and B. R. Webber, “Associated jet and subjet rates in light-quark and gluon jet discrimination,” JHEP **1504** (2015) 131.
- [11] A. J. Larkoski, J. Thaler and W. J. Waalewijn, “Gaining (Mutual) Information about Quark/Gluon Discrimination,” JHEP **1411** (2014) 129.
- [12] Y. Sakaki, “Evolution variable dependence of jet substructure,” JHEP **1508** (2015) 100.

- [13] J. R. Andersen *et al.*, “Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report,” arXiv:1605.04692 [hep-ph].
- [14] D. Ferreira de Lima, P. Petrov, D. Soper and M. Spannowsky, “Quark-Gluon tagging with Shower Deconstruction: Unearthing dark matter and Higgs couplings,” arXiv:1607.06031 [hep-ph].
- [15] D. Goncalves, F. Krauss and R. Linten, “Distinguishing b-quark and gluon jets with a tagged b-hadron,” Phys. Rev. D **93** (2016) no.5, 053013.
- [16] I. Moutl, L. Necib and J. Thaler, “New Angles on Energy Correlation Functions,” arXiv:1609.07483 [hep-ph].
- [17] G. Aad *et al.* [ATLAS Collaboration], “Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector,” Eur. Phys. J. C **74** (2014) 8, 3023.
- [18] CMS Collaboration, “Performance of quark/gluon discrimination using pp collision data at $\sqrt{s} = 8$ TeV,” CMS-PAS-JME-13-002.
- [19] The ATLAS collaboration [ATLAS Collaboration], “Discrimination of Light Quark and Gluon Jets in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector,” ATLAS-CONF-2016-034.
- [20] T. Sjostrand, S. Mrenna and P. Z. Skands, “PYTHIA 6.4 Physics and Manual,” JHEP **0605**, 026 (2006).
- [21] T. Sjostrand, S. Mrenna and P. Z. Skands, “A Brief Introduction to PYTHIA 8.1,” Comput. Phys. Commun. **178** (2008) 852; T. Sjostrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna and S. Prestel *et al.*, “An Introduction to PYTHIA 8.2,” Comput. Phys. Commun. **191** (2015) 159
- [22] M. Bahr, S. Gieseke, M. A. Gigg, D. Grellscheid, K. Hamilton, O. Latunde-Dada, S. Platzer and P. Richardson *et al.*, “Herwig++ Physics and Manual,” Eur. Phys. J. C **58** (2008) 639.
- [23] G. Aad *et al.* [ATLAS Collaboration], “Measurement of the charged-particle multiplicity inside jets from $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector,” Eur. Phys. J. C **76** (2016) no.6, 322
- [24] The ATLAS collaboration, “Search for Supersymmetry at the high luminosity LHC with the ATLAS Detector”, ATL-PHYS-PUB-2014-010.
- [25] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer and T. Stelzer, “MadGraph 5 : Going Beyond,” JHEP **1106**, 128 (2011);
J. Alwall *et al.*, “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations,” JHEP **1407** (2014) 079.
- [26] C. Borschensky, M. Kraemer, A. Kulesza, M. Mangano, S. Padhi, T. Plehn and X. Portell, “Squark and gluino production cross sections in pp colli-

- sions at $\sqrt{s} = 13, 14, 33$ and 100 TeV,” Eur. Phys. J. C **74** (2014) no.12, 3174; See also, <https://twiki.cern.ch/twiki/bin/view/LHCPhysics/SUSYCrossSections14TeVgluglu>
- [27] J. Pumplin, D. R. Stump, J. Huston, H. L. Lai, P. M. Nadolsky and W. K. Tung, “New generation of parton distributions with uncertainties from global QCD analysis,” JHEP **0207** (2002) 012.
 - [28] M. R. Whalley, D. Bourilkov and R. C. Group, “The Les Houches accord PDFs (LHAPDF) and LHAGLUE,” hep-ph/0508110.
 - [29] S. Ovnyn, X. Rouby and V. Lemaitre, “DELPHES, a framework for fast simulation of a generic collider experiment,” arXiv:0903.2225 [hep-ph]; J. de Favereau *et al.* [DELPHES 3 Collaboration], “DELPHES 3, A modular framework for fast simulation of a generic collider experiment,” JHEP **1402** (2014) 057.
 - [30] M. Cacciari, G. P. Salam and G. Soyez, “FastJet user manual,” Eur. Phys. J. C **72** (2012) 1896; M. Cacciari and G. P. Salam, “Dispelling the N^3 myth for the k_t jet-finder,” Phys. Lett. B **641** (2006) 57.
 - [31] M. Cacciari, G. P. Salam and G. Soyez, “The Anti- $k(t)$ jet clustering algorithm,” JHEP **0804** (2008) 063.
 - [32] A. Hocker, J. Stelzer, F. Tegenfeldt, H. Voss, K. Voss, A. Christov, S. Henrot-Versille and M. Jachowski *et al.*, “TMVA - Toolkit for Multivariate Data Analysis,” PoS ACAT (2007) 040 [physics/0703039 [PHYSICS]]; P. Speckmayer, A. Hocker, J. Stelzer and H. Voss, “The toolkit for multivariate data analysis, TMVA 4,” J. Phys. Conf. Ser. **219** (2010) 032057; <http://tmva.sourceforge.net>
 - [33] Rene Brun and Fons Rademakers, "ROOT - An Object Oriented Data Analysis Framework", Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86. See also <http://root.cern.ch/>;